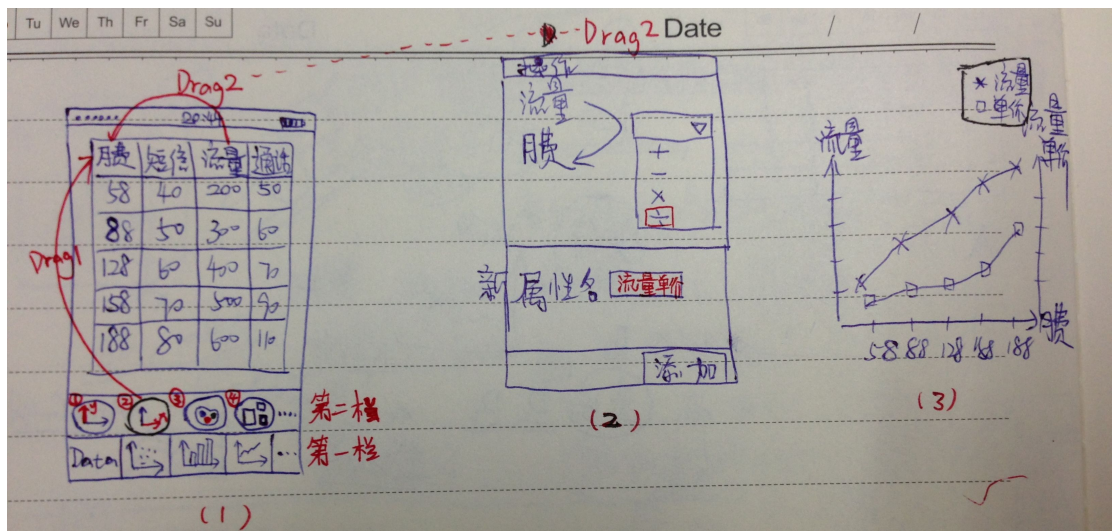


Weekly Report (2014.09.22~09.28)

Done

1) 手机拍照可视化的项目，目前的策略是：

- 照片-->数据转换部分暂时不做，主要是由于目前我们能利用的 OCR 相关资源效果都不好，唯一效果还不错的 Office Lens 暂时没有可用 API。由于这部分只是涉及一些图像识别的技术问题，不涉及到可视化的本身，对项目的影响也不大。
- 数据-->可视化部分：最终的可视化部分，我们只需要支持最基本的比如直方图、散点图、折线图可视化形式。重点放在如何将照片识别的结果作用于可视化，即通过简单的交互，帮助用户将数据映射到可视化编码的各个通道上。我的初步设想如下：
 - 主要视图如图（1）所示，最下面一栏的第一项“data”被选中，此时展示数据表格视图，以及第二栏工具栏。
 - Drag1: 用户可以拖动第二栏的图标到数据的任意维度对数据进行绑定：比如 Drag1 所示，将“x 轴”拖到“月费”，表示将“月费”绑定到“x 轴”上。同理，可以绑定 y 轴，颜色，尺寸等维度。
 - Drag2: 用户还可以将任意维度拖动到其余维度上，做简单的代数运算，并生成新的一列数据维度。比如图（2）中将“流量”拖动到“月费”，并选中“÷”，定义新属性名为“流量单价”，将生成新的一列，定义流量的单价。
 - 当用户在绑定好数据后单击底部工具栏的其他图标（散点图、直方图或折线图图标），将直接展示可视化的结果。
 - 此外，图（1）中还应包括对 OCR 操作识别错误的的数据（或属性名）进行修改的操作。



可视化展示部分，我准备试试看 baidu 的 ECharts，包括数据的指定，也可以拿 ECharts 做参考。

2) 关于时空数据检索的项目，目前有以下两个方案，但是两个方案的优劣，还需进行实现并比较。这两个方案首先均在时空维度上进行细分，形成多个 time-space cell，每个 cell 存储了相应时空相关的数据索引。在每个 cell 索引的数据的存储方式上，两个方案有所不同（以下描述以人的移动数据为例）：

- 第一个方案是常见的方法：将 cell 内的数据按 hash 表的方式存，直接对人的 ID 进行 hash。在多个 cell 做“交”操作的时候，我们首先选择出 size 最小的 cell，然后对于

这个 cell 里面的每个 key，在其他 cell 中一一查询是否存在，如果不存在，则将这条数据删去，最后剩下的数据就是我们要的结果。

- b) 第二个方案将 PSH 方法融入到 cell 里面：假设有 1000,000 人，那么每个 cell 里面开辟一个 1000×1000 的数组，数组里面每个元素对应一个人，1 表示这个人出现在这个 time-space cube 里面，0 则表示这个人没有出现。然后用 PSH 对这个数组进行压缩表示。在多个 cell 做“交”操作的时候，我们将每个 cell 的 PSH 压缩表示作为纹理，加载到显存中，直接做“按位与”操作，则最终得到的 1000×1000 数组里面值为 1 的元素即为我们需要的结果。

由于对 GPU 操作的细节我目前还不是很清楚，暂时不能直接判断孰优孰劣。海东的意见是，即便将 PSH 的压缩表示用 GPU 进行“按位与”操作，由于需要逐位解压求交，效率并不见得会快。

对于方案二，假设空间分为 1000×1000 份，时间按分钟将一天分为 1440 份，每个索引占用 4Byte，稀疏率为 5%，那么对于一天的数据，索引占用的最小空间为 $1000 \times 1000 \times 1440 \times 4\text{Byte} \times 5\% = 288\text{M}$ ，索引数据占用的最小空间为 $1000,000 \times 1440 / 8 = 180\text{M}$ 。一共为 468M。

To Do

- 1) 参考 ECharts，将数据修正、指定的过程进行细化。
- 2) 对于时空数据检索方面，将两种方案进行实现，数据采用汪飞之前使用的 1 天的出租车数据。